4th 第四届博鳌国际药械真实世界研究大会
BOAO INTERNATIONAL CONFERENCE ON REAL WORLD STUDIES OF MEDICAL PRODUCTS

大型前列腺癌临床数据集向 OMOP 通用数据模型的转换 – 基于中国区域医疗数据库的经验

Transforming a Large-Scale Prostate Cancer Outcomes Dataset to the OMOP Common Data Model - Experiences with a Regional Health Record Database in China

汇报人：许哲 **单位**：拜耳医药保健有限公司
Presenter: Zhe Xu **Affiliation**: Bayer Healthcare Company Ltd
代表项目作者汇报 On behalf of co-authors

许哲 Zhe Xu

拜耳医药保健有限公司
Bayer Healthcare Company Ltd.

作者：蔡启亮[1], 谢洋[2],李静[2],林晓羽[2],杨芷其[2],殷铮[2],叶怡君[2],黄诗韵[3], 许哲[3], 牛远杰[4]
[1]天津医科大学第二医院泌尿外科，天津，中国
[2]艾昆纬企业管理咨询有限公司真实世界研究，上海，中国
[3]拜耳医药保健有限公司医学事务部，北京，中国
[4]天津医科大学总医院泌尿外科，天津，中国

Authors: Qiliang Cai[1], Yang Xie[2], Jing Li[2], Xiaoyu Lin[2], Chih-Chi Yang[2], Zheng Yin[2], Yi-Chun Yeh[2], Shiyun Huang[3], Zhe Xu[3], Yuanjie Niu[4]
[1]Department of Urology, Second Hospital of Tianjin Medical University, Tianjin, China.
[2]Real World Solutions, IQVIA Solutions Enterprise Management Consulting (Shanghai) Co., Ltd, Shanghai, China
[3]Medical Affairs, Pharmaceuticals, Bayer Healthcare Company Ltd, Beijing, China.
[4]Department of Urology, Tianjin Medical University General Hospital, Tianjin, China.

- 讲者简介 Speaker Bio

许哲，博士，现任拜耳医药保健有限公司医学部证据生成高级经理，负责多个产品的证据生成计划制定，专病数据库搭建及数据标准化项目，真实世界研究等。
先后于北京大学、美国约翰霍普金斯大学、英国剑桥大学获得预防医学学士、流行病学硕士、博士学位，主要研究方向为基于大规模电子病历数据的心血管疾病风险预测模型。

Zhe Xu, PhD. I currently work as the Senior Manager of Evidence Generation at Bayer HealthCare Co. Ltd., mainly responsible for the evidence generation plans, establishment of real-world data cohorts, database standardization, and real-world studies, for multiple products.

I have obtained a Bachelor's degree in Preventive Medicine from Peking University, a Master's degree in Epidemiology from Johns Hopkins University, and a Ph.D. in Epidemiology from the University of Cambridge. My main research focus is on cardiovascular disease risk prediction models based on large-scale electronic health record data.

由OHDSI维护的OMOP-CDM为真实世界数据分析提供一个标准化框架

OMOP-CDM, maintained by OHDSI, provides a standardized framework for real-world data analysis

实现快速且可靠的研究 Enables rapid and reliable research

标准化使研究人员能够在各种数据集和数据类型中快速而可靠地开展研究

Standardization allows researchers to conduct reliable studies across a variety of datasets and data types.

降低维护成本 Reduces the cost of ownership

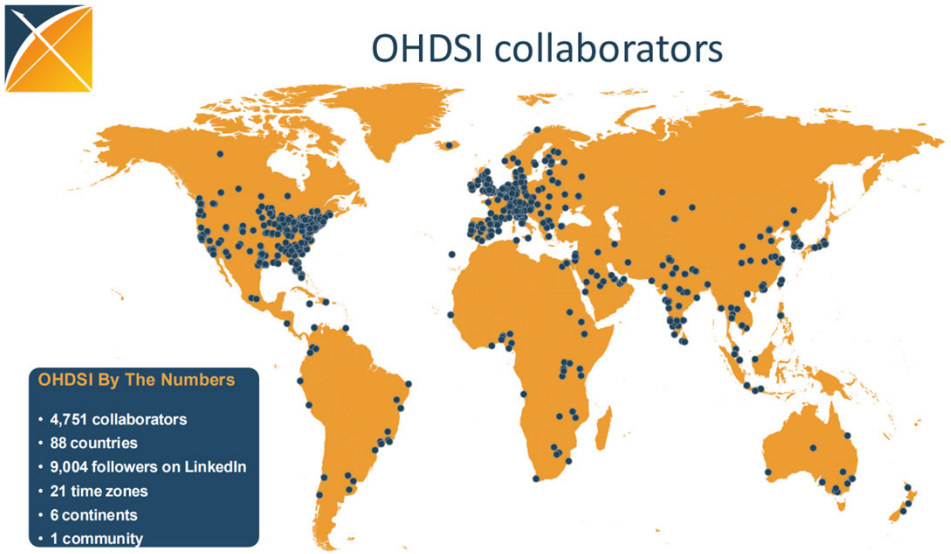
通过减少理解数据模式和编写跨数据库分析代码所需的工作量来降低成本

Standardization lowers costs by reducing the effort required to understand data schemas and to write cross-database analytical code.

扩展数据访问 Expands data accessibility

OMOP-CDM 借助 OHDSI 网络的分布式多中心研究实现更广泛的数据访问

The OMOP-CDM facilitates expanded data access via the OHDSI network and remote multi-site database studies.



Abbreviation: OHDSI = Observational Health Data Sciences and Informatics; OMOP-CDM = Observational Medical Outcomes Partnership Common Data Model

Reference:

OHDSI Collaborators. Accessed from https://www.ohdsi.org/who-we-are/collaborators/

OMOP 被全球各国政府机构采用
OMOP is adopted by government agencies globally

国家药品监督管理局 (NMPA) 的指南中引用了通用数据模型 (CDM)

The Common Data Model (CDM) is cited in the guideline from the National Medical Products Administration (NMPA)

- ***《用于产生真实世界证据的真实世界数据指导原则（试行）》***
  - 2021年4月15日发布
  - 由国家药品监督管理局药品审评中心 (CDE) 发布
  - 自发布之日施行

- ***Guidance for Real-World Data Used to Generate Real-World Evidence (Trial Implementation)***
  - Date of Issuance: April 15, 2021
  - Issuing Authority: Center for Drug Evaluation (CDE) of the National Medical Products Administration (NMPA)
  - Effective date: Effective immediately upon publication (April 15, 2021)

国家药品监督管理局药品审评中心
CENTER FOR DRUG EVALUATION, NMPA

当前位置：新闻中心>>工作动态>>通知公告>>新闻正文

国家药监局药审中心关于发布《用于产生真实世界证据的真实世界数据指导原则（试行）》的通告（2021年第27号）

发布日期：20210415

为进一步指导和规范申办者利用真实世界数据生成真实世界证据支持药物研发，药审中心组织制定了《用于产生真实世界证据的真实世界数据指导原则（试行）》（见附件），根据《国家药监局综合司关于印发药品技术指导原则发布程序的通知》（药监综药管〔2020〕9号）要求，经国家药品监督管理局审查同意，现予发布，自发布之日起施行。

特此通告。

国家药品监督管理局药品审评中心

2021年4月13日

附件1：《用于产生真实世界证据的真实世界数据指导原则（试行）》.pdf

Abbreviation: NMPA = National Medical Products Administration; CDE = Center for Drug Evaluation

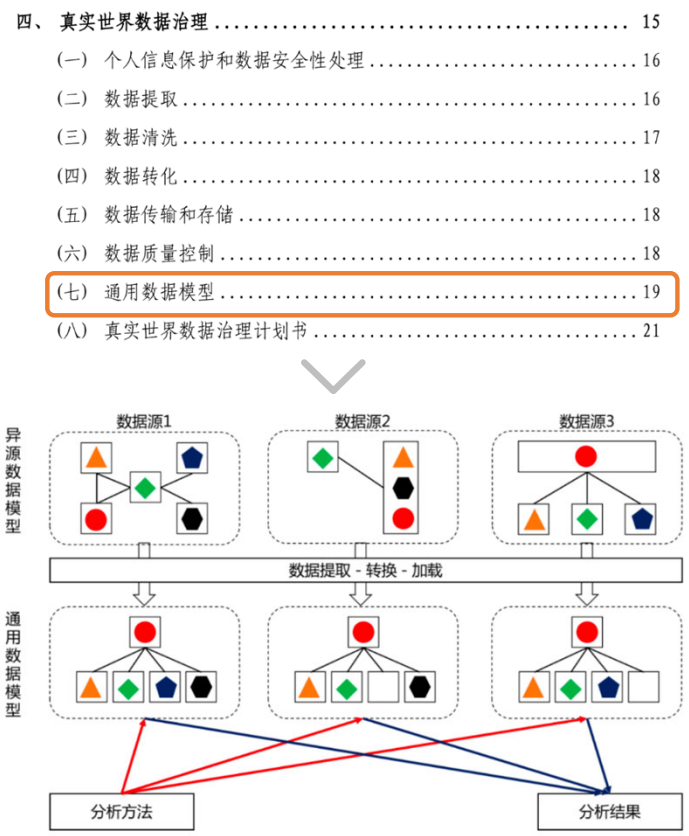
Reference:
1. Guidance for Real-World Data Used to Generate Real-World Evidence (Trial Implementation)
Accessed from https://www.cde.org.cn/main/news/viewInfoCommon/2a1c437ed54e7b838a7e86f4ac21c539.

国家药品监督管理局 (NMPA) 的指南中引用了通用数据模型 (CDM)

The Common Data Model (CDM) is cited in the guideline from the National Medical Products Administration (NMPA)

“由于多源数据的结构和类型的复杂性、样本规模和标准的差异性，在将源数据转换为通用数据模型的整体过程中，需要对源数据进行提取、转换、加载，应确保源数据在语法和语义上与目标分析数据库的结构和术语一致。[1].

Given the complexity of multi-source data structures and types, and the heterogeneity in sample sizes and standards, converting source data into a common data model requires extraction, transformation, and loading. Source data must be syntactically and semantically aligned with the structure and terminology of the target analytical database[1].”



Abbreviation: NMPA = National Medical Products Administration; CDE = Center for Drug Evaluation

Reference:
1. Guidance for Real-World Data Used to Generate Real-World Evidence (Trial Implementation) Accessed from https://www.cde.org.cn/main/news/viewInfoCommon/2a1c437ed54e7b838a7e86f4ac21c539.

OMOP 数据源
OMOP data source

数据来源

- 截至2024年，天津电子健康档案数据库 (rEHR) 包含来自公立医疗机构的超过1100亿条记录，其中290亿条记录已通过数据治理处理，覆盖自1995年以来超过1600万居民。
- 数据整合自**43家三级医院**和**39家二级医院**，以及公共卫生记录，覆盖自2000年以来约1200万人。
- 项目**提取了2017年至2021年的数据**，并转换为 OMOP-CDM 模型，研究级数据质量自2015年起可用。

Data source

- As of 2024, the Tianjin regional electronic health record database (rEHR) includes over 110 billion records from public medical institutions, with 29 billion records processed through data governance, covering over 16 million residents since 1995.
- It integrates data from **43 tertiary hospitals** and **39 secondary hospitals**, along with public health records, covering ~12 million people since 2000.
- For this study, **data from 2017 to 2021** were extracted and transformed into the OMOP-CDM schema, with research-grade data quality available since 2015.

Abbreviation: rEHR = regional Electronic Health Record database; OMOP-CDM = Observational Medical Outcomes Partnership Common Data Model

队列定义
Cohort definitions

样本选择

- 纳入的患者在**天津电子健康档案数据库 (rEHR)** 的门诊或住院记录中**至少有一次前列腺癌 (PC) 诊断。**
- PC诊断通过ICD-10代码C61或诊断名称包含“前列腺”，并伴随“癌”、“原位癌”、“CA”、“恶性肿瘤”、“PCA”、“CRPC”或“HSPC”等术语进行识别。
- 如诊断名称包含“OPCA”、“PCAD”、“前列腺增生”或“肾癌”等术语，则排除，以确保病例识别的准确性。

Sample selection

- Included **patients had at least one diagnosis of Prostate Cancer** (PC) recorded during outpatient visits or hospitalizations in the **Tianjin rEHR database**.
- PC diagnoses were identified using ICD-10 code C61 or diagnosis names containing “前列腺” along with terms like “癌”, “原位癌”, “CA”, “恶性肿瘤”, “PCA”, “CRPC”, or “HSPC”.
- Records were excluded if diagnosis names included terms such as “OPCA”, “PCAD”, “前列腺增生”, or “肾癌” to ensure accurate case identification.

Abbreviation: rEHR = regional Electronic Health Record database; OMOP-CDM = Observational Medical Outcomes Partnership Common Data Model; PC = Prostate Cancer; ICD-10 = The International Statistical Classification of Diseases and Related Health Problems 10th Revision; CA = Cancer; PCA = Prostate Cancer; CRPC = Castration-Resistant Prostate Cancer; HSPC = Hormone-Sensitive Prostate Cancer; OPCA = Occult Prostate Cancer; PCAD = Prostate Cancer Associated Disease

OMOP-CDM 转换工作流程
OMOP-CDM conversion workflow

1 源数据探查 Source Data Profiling

- 分析源数据的结构和内容，以了解其格式、完整性和质量。
- Analyze the structure and content of the source data to understand its format, completeness, and quality.

2 ETL 映射到 CDM模式 ETL Mapping to CDM

- 使用提取、转换和加载 (ETL) 操作将源数据元素映射到 CDM 模式的指定目标位置。
- Map source data elements to the CDM schema using extract, transform, and load(ETL) operations.

3 术语标准化 Vocabulary Standardization

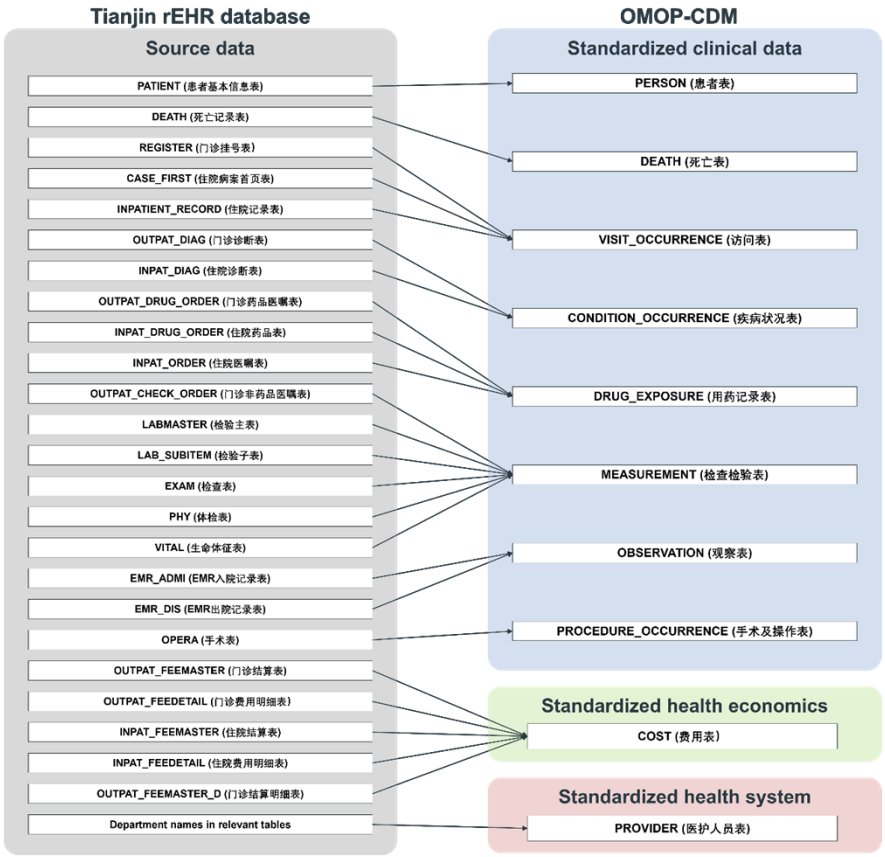
- 将源数据中使用的代码和值转换为 CDM 中采用的代码和值，以实现数据的一致性与互操作性。
- Translate source codes and values into standardized CDM vocabularies for consistency and interoperability.

4 质量保证 Quality Assurance

- 执行验证和质量检查，以确保数据的准确性、完整性以及符合 CDM 标准。
- Perform validation and quality checks to ensure data accuracy, completeness, and conformance to CDM standards.

Abbreviation: ETL = Extract, Transform, Load; CDM = Common Data Mode

将天津电子健康档案数据库转换为OMOP-CDM
Transformation of the Tianjin rEHR Database into OMOP-CDM



天津电子健康档案数据已成功映射到 OMOP-CDM 标准化表格，确保后续分析和研究的互操作性和一致性。
The Tianjin rEHR data have been successfully mapped to OMOP-CDM standardized tables, ensuring interoperability and consistency for downstream analytics and research.

天津电子健康档案数据库 (rEHR) 提供 OMOP-CDM 所需的**所有**临床信息，涵盖**三个**主要类别:
The Tianjin rEHR database provides **all** clinical information required for OMOP-CDM, covering **three** major categories:

- 临床数据 Clinical Data
- 健康经济学 Health Economics
- 卫生系统数据 Health System Data

Abbreviation: rEHR = regional Electronic Health Record database; OMOP-CDM = Observational Medical Outcomes Partnership Common Data Model; EMR = Electronic Medical Record

天津电子健康档案数据库成功完成OMOP-CDM映射和质量控制
Successful OMOP-CDM Mapping and Quality Assurance in Tianjin rEHR Database

领域 Domain	映射逻辑 Mapping logic	映射率 (记录层级) Mapping rate (records level)
诊断 (住院+门诊) Diagnosis (inpatient + outpatient)	• 原始诊断名称首先通过中文名称匹配映射至 ICD10CN 编码。 • 对于包含多个诊断的原始诊断，在适用的情况下将其映射至多个类型域。 • 所有已映射的诊断名称最终将统一标准化为 OMOP 的 condition 类型域。 • Diagnosis names mapped to ICD10CN and OMOP concepts. • Multiple diagnoses mapped to multiple concepts as needed. • Standardized to OMOP concepts.	89%
用药 (住院+门诊) Drug (inpatient + outpatient)	• 原始药品名称通过中文名称匹配映射至 ATC 编码 • 所有已映射的药物名称最终将统一标准化为 OMOP 的drug ingredient类型域。 • Drug names mapped to ATC and OMOP concepts. • Standardized to OMOP drug ingredients.	85%
手术 Procedure	• 原始手术名称通过中文名称匹配映射至 ICD9ProcCN 编码；对于未能匹配的名称，在可行的情况下进一步映射至 OMOP 概念。 • 所有已映射的手术名称最终将统一标准化为 OMOP 的 procedure 类型域。 • Operation names mapped to ICD9ProcCN via Chinese name matching; remaining names mapped to OMOP where possible. • Standardized to OMOP procedure concepts.	89%
实验室检测 Lab tests	• 原始实验室检测名称首先通过中文名称匹配映射至 LOINC CN 编码；对于未能匹配的名称，在可行的情况下进一步映射至 OMOP 概念。 • 所有已映射的实验室检测名称最终将统一标准化为OMOP的measurement类型域。 • Lab names mapped to LOINC CN via Chinese name matching; remaining names mapped to OMOP where possible. • Standardized to OMOP measurement concepts.	87%
实验室检测单位 Lab test units	• 原始实验室检测单位将直接映射至 OMOP 的 unit 类型域。 • Raw lab test units were mapped directly to OMOP unit concepts.	98%

天津电子健康档案数据库成功完成OMOP-CDM映射和质量控制
Successful OMOP-CDM Mapping and Quality Assurance in Tianjin rEHR Database

领域 Domain	映射率 (记录层级) Mapping rate (records level)
诊断 (住院+门诊) Diagnosis (inpatient + outpatient)	89%
用药 (住院+门诊) Drug (inpatient + outpatient)	85%
手术 Procedure	89%
实验室检测 Lab tests	87%
实验室检测单位 Lab test units	98%

天津电子健康档案数据库 (rEHR) **超过85%的临床数据**已成功映射至 OMOP-CDM 结构
Over 85% of clinical records in the Tianjin rEHR database were successfully mapped to the OMOP-CDM structure.

转换后的数据库通过**126项质量检查**，涵盖以下领域的验证：
The transformed database **passed all 126 quality checks**, including validation of the following domains:

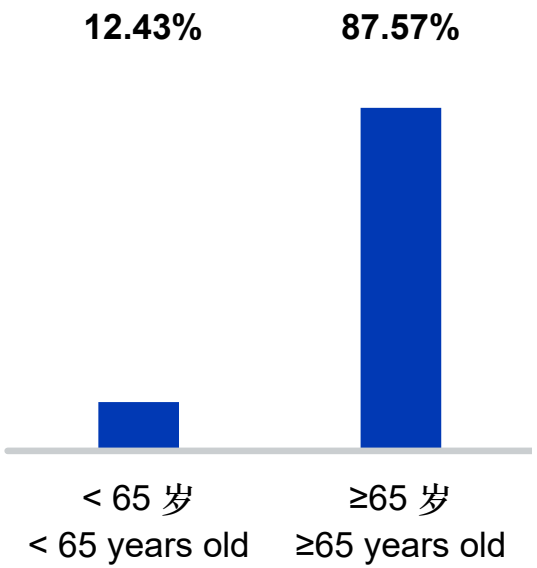
- 患者信息 Patient information
- 诊断 Diagnoses
- 药物 Medications
- 手术 Procedures
- 实验室检测 Laboratory measurements

Abbreviation: rEHR = regional Electronic Health Record database; OMOP-CDM = Observational Medical Outcomes Partnership Common Data Model

前列腺癌患者的特征分析
Characteristics of prostate cancer patients

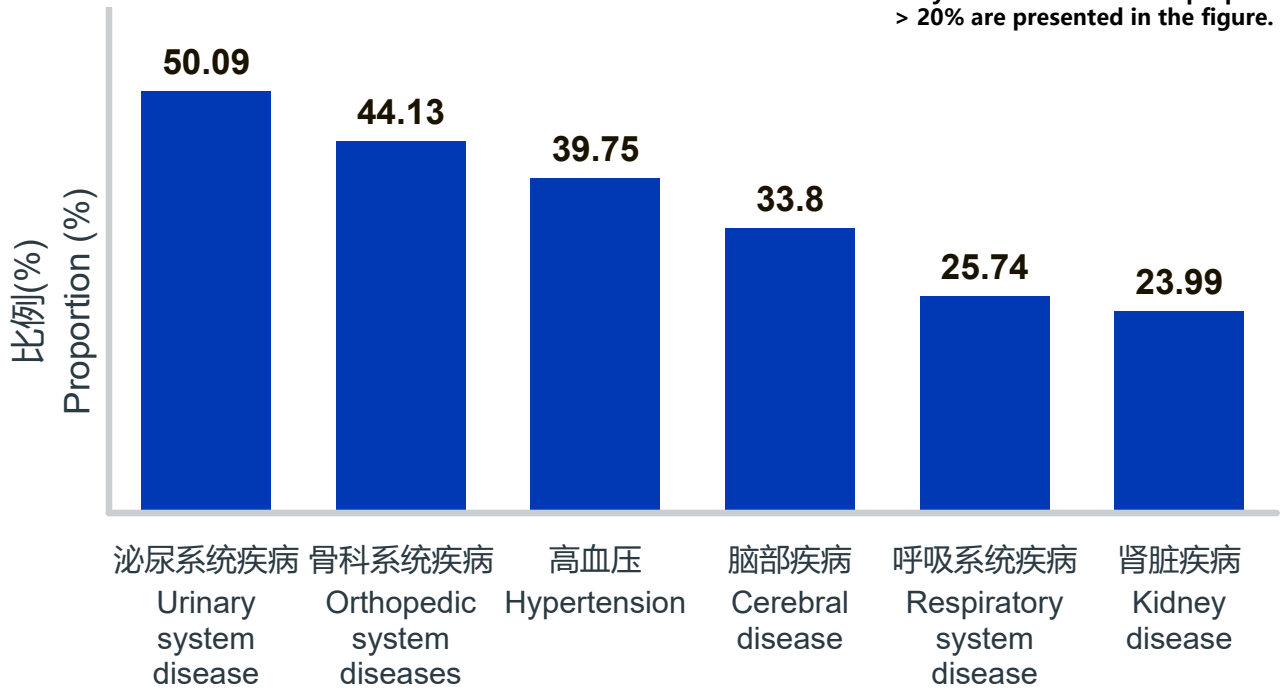
4th BOAO INTERNATIONAL CONFERENCE ON REAL WORLD STUDIES OF MEDICAL PRODUCTS
第四届博鳌国际药械真实世界研究大会

诊断时的年龄组
Age group at diagnosis

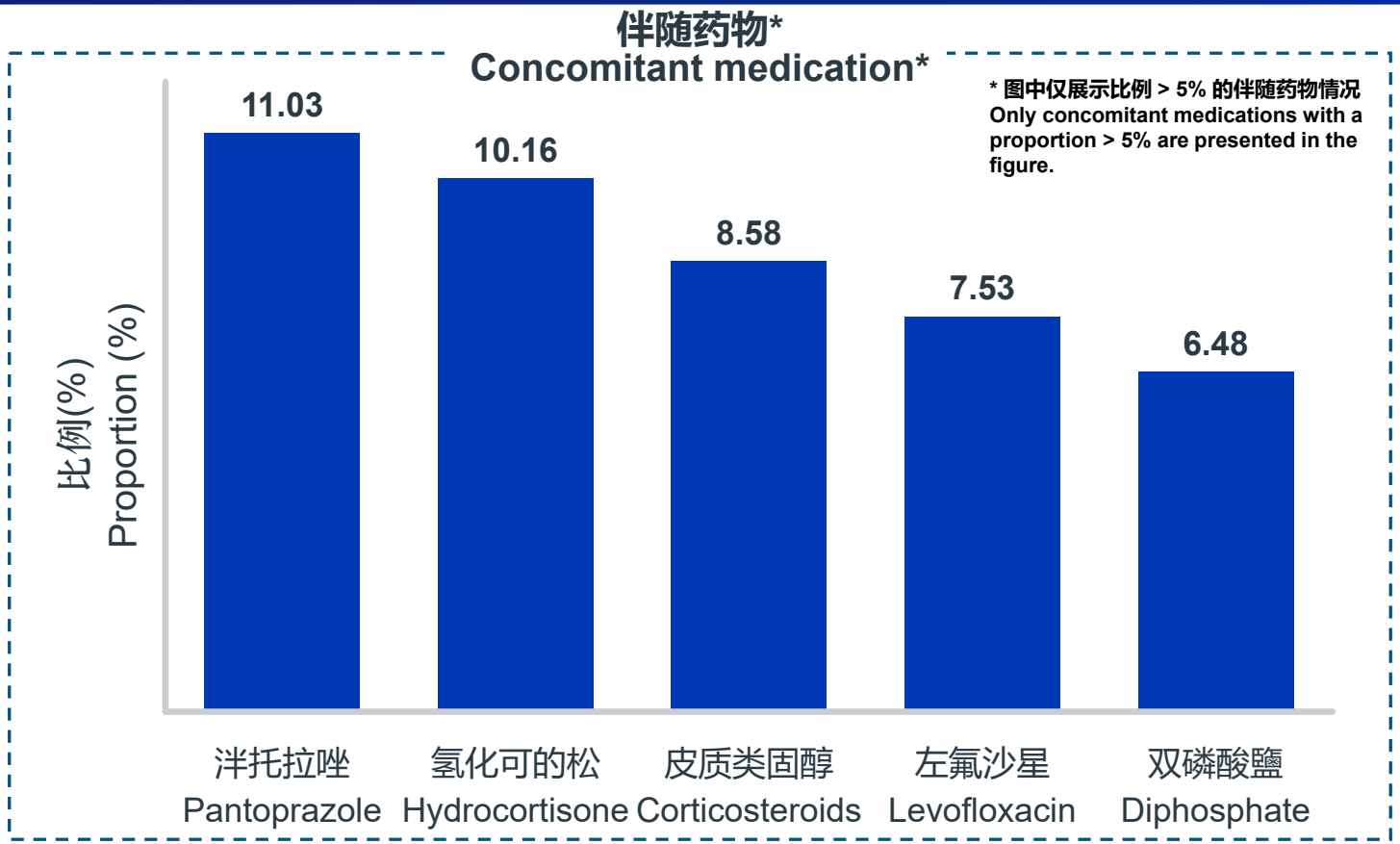


共病*
Comorbidity*

* 图中仅展示比例 > 20% 的共病情况
Only comorbidities with a proportion > 20% are presented in the figure.



前列腺癌患者的特征分析
Characteristics of prostate cancer patients



结论
Conclusion

天津电子健康档案数据库的前列腺癌队列已成功转换为 OMOP-CDM
The prostate cancer cohort of the Tianjin rEHR database has been successfully converted to the OMOP-CDM

本研究表明中国的区域电子健康档案数据可在**信息损失最小**的情况下实现标准化
This study demonstrated that standardizing regional EHR data in China is feasible with **minimal information loss**.

此次数据转换凸显了 **OMOP-CDM** 在中国的其他区域电子健康档案数据库或独立电子病历数据库中**应用的巨大潜力**
This success highlights the **potential to scale the OMOP-CDM approach** to other regional or standalone EMR systems across China.

这一标准化工作有助于**提升数据互操作性**，符合国家健康数据整合的目标，并进一步**强化了在本研究中应用该框架的价值**
This standardization effort supports **enhanced data interoperability**, aligns with national health data integration goals, and **reinforces the value of applying this framework in the current study**.

Abbreviation: rEHR = regional Electronic Health Record database; OMOP-CDM = Observational Medical Outcomes Partnership Common Data Model; EMR = Electronic Medical Record

4th 第四届博鳌国际药械真实世界研究大会

BOAO INTERNATIONAL CONFERENCE ON REAL WORLD STUDIES OF MEDICAL PRODUCTS

感谢聆听!
Thank you!

姓名：许哲 **单位**：拜耳医药保健有限公司

Presenter: Zhe Xu **Affiliation**: Bayer Healthcare Company Ltd